Hierarchical Bayesian modelling of the electricity load

Tristan Launay, Anne Philippe, Sophie Lamarche

Hierarchical Bayesian modelling of the electricity load

Tristan Launay[1,2] Anne Philippe[1] Sophie Lamarche[2]

[1] Laboratoire de Mathématiques Jean Leray, 2 Rue de la Houssinière – BP 92208, 44322 Nantes Cedex 3, France

[2] Electricité de France R&D, 1 Avenue du Général de Gaulle, 92141 Clamart Cedex, France

Abstract

In this paper, we study a non-linear model used to estimate and forecast the electricity load, that usually requires four or more years worth of data to avoid any overfitting phenomenon. We first propose a non-informative prior to be used when the number of observations is large enough. When the observations are too few, we propose a hierarchical prior to include information coming from another bigger, similar, sample. The posterior densities associated with these two priors are derived and a MCMC algorithm is provided in each case. We finally run these algorithms on simulated and real datasets ; the hierarchical prior greatly improves the quality of the model predictions.

keywords : informative prior ; mcmc algorithms ; small dataset ; electricity load forecasting

1 Introduction

Modelling and forecasting the electricity load (or demand) on a day-to-day basis has long been a key activity for any company involved in the electricity industry. It is first and foremost needed to supply a fixed voltage at all ends of an electricity grid : to be able to do so, the amount of electricity produced has to match the demand very closely at any given time and experts usually make use of short-term forecasts with this aim in view as mentionned in Cottet and Smith (2003). However long-term forecasts are also required when it comes down to effectively scheduling maintenance operations over the network, whose production units range from nuclear powerplants to wind turbines.

The advent of the wholesale electricity market in Europe and in France (since 2003 for industry, since July 2007 for every customer) has brought renewed focus on load forecasting for different reasons. For instance, a small improvement of the load forecasts can sometimes lead to important financial benefits, especially during peak load periods when prices happen to reach very high levels.

Electricity load usually has a large predictable component due to its very strong daily, weekly and yearly periodic behaviour. It has also been noted in many regions that the weather usually affects the load too, the most important meteorological factor typically being the temperature (see Al-Zayer and Al-Ibrahim, 1996, for an example).

The EVENTAIL model (see Bruhns et al., 2005), used to describe and forecast the electricity load is a multi-equations non-linear regression model : each of the 48 instants of the day (each one lasts 30 minutes, starting at 00:00AM) is modelled using a separate parametric regression which makes it possible for the parameters to actually depend on the time of day. Since all instants are treated in the same way, the paper focuses on one only. For a given instant of the day (10:00AM, say), the model is made of three components, which we explain briefly in the next paragraphs,

and is usually formulated as follows : for $t = 1, \ldots, N$,

$$
\begin{aligned}
y_t &= x_t^{(1)} \cdot x_t^{(2)} + x_t^{(3)} + \epsilon_t \qquad (1)\\
x_t^{(1)} &= \sum_{j=1}^{d_{11}} \left[z_j^{\cos} \cos\left(\frac{2j\pi}{365.25} \times t \right) + z_j^{\sin} \sin\left(\frac{2j\pi}{365.25} \times t \right) \right] + \sum_{j=1}^{d_{12}} \omega_j \mathbb{1}_{\Omega_j}(t),\\
x_t^{(2)} &= \sum_{j=1}^{d_2} \psi_j \mathbb{1}_{\Psi_j}(t),\\
x_t^{(3)} &= g(T_t - u)\mathbb{1}_{[T_t, +\infty[}(u),
\end{aligned}
$$

where y_t is the load of day t and where $\epsilon_1, \ldots, \epsilon_N$ are assumed independent and identically distributed with common distribution $\mathcal{N}(0, \sigma^2)$[1].

The $x^{(1)}$ component is meant to account for the average seasonal behaviour of the electricity load, with a truncated Fourier series (whose coefficients are $z_j^{\cos} \in \mathbb{R}$ et $z_j^{\sin} \in \mathbb{R}$) and gaps (parameters $\omega_j \in \mathbb{R}$) which represent the average levels of electricity load over predetermined periods given by a partition $(\Omega_j)_{j \in \{1,\ldots,d_{12}\}}$ of the calendar. This partition usually specifies holidays, or the period of time when daylight saving time is in effect i.e. major breaks in the electricity consumption behaviour. Figure 1 shows a typical behaviour over two different periods of time (summer vs. winter).

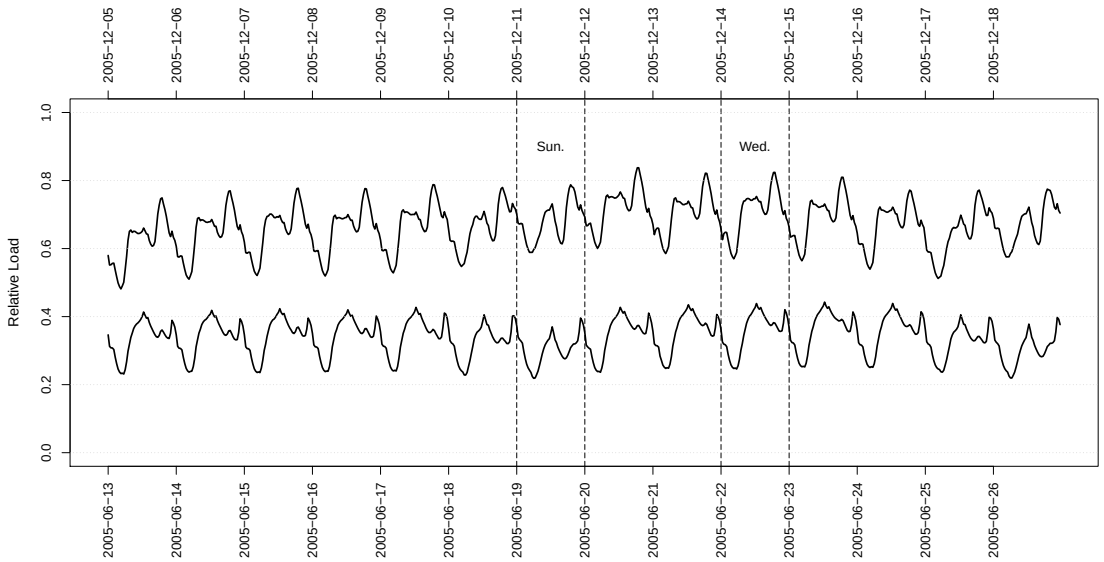

Figure 1: In black : relative electricity load over 2 June weeks from 13/06/2005 to 26/06/2005 for a population. In grey : relative electricity load over 2 December weeks from 05/12/2005 to 18/12/2005 for the same population. Relative load, shown on the y-axis of the graph, is the ratio between the electricity load and its maximum value. Two days have been highlighted to show the difference between weekdays and weekends. Also notice the daily patterns of the electricity load are not the same during summer and during winter.

The $x^{(2)}$ component allows for day-to-day adjustments of the seasonal behaviour $x^{(1)}$ through shapes (parameters ψ_j) that depends on the so-called days' types which are given by a second partition $(\Psi_j)_{j \in \{1,\ldots,d_2\}}$ of the calendar. This partition usually separates weekdays from weekends, and bank holidays. The differences between two different daytypes are visible on Figure 1 too. For

[1] $\mathcal{N}(m, \Sigma)$ is the Gaussian distribution with density $(2\pi)^{-d/2}|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-m)'\Sigma^{-1}(x-m)\right)$ in which $'$ denotes the transposition operator.

obvious identifiability reasons, the vector ψ is restricted to the positive quadrant of the $\|\cdot\|_1$-unit sphere in $\mathbb{R}^{d_2}$, that we denote

$$S_+^{d_2}(0,1) = \left\{\psi \in (\mathbb{R}_+)^{d_2}, \quad \|\psi\|_1 = 1\right\}.$$

The $x^{(3)}$ component represents the non-linear heating effect that links the electricity load to the temperature, with the help of 2 parameters. The heating threshold $u \in [\underline{u}, \overline{u}]$ corresponds to the temperature above which the heating effect is considered null and is usually estimated to be roughly around 15°C. The heating effect is supposed to be linear for temperatures below the threshold and null for temperatures above. The restriction on the support of the threshold u simply expresses the fact that the threshold is sought within the range of the observed temperatures, i.e. $u \in [\underline{u}, \overline{u}]$ with

$$\min_{t=1,\dots,N} T_t < \underline{u} < \overline{u} < \max_{t=1,\dots,N} T_t.$$

The heating gradient $g \in \mathbb{R}^*$ represents the intensity of the heating effect, i.e. the slope (assumed to be non-zero) of the linear part that can be observed on figure 2.

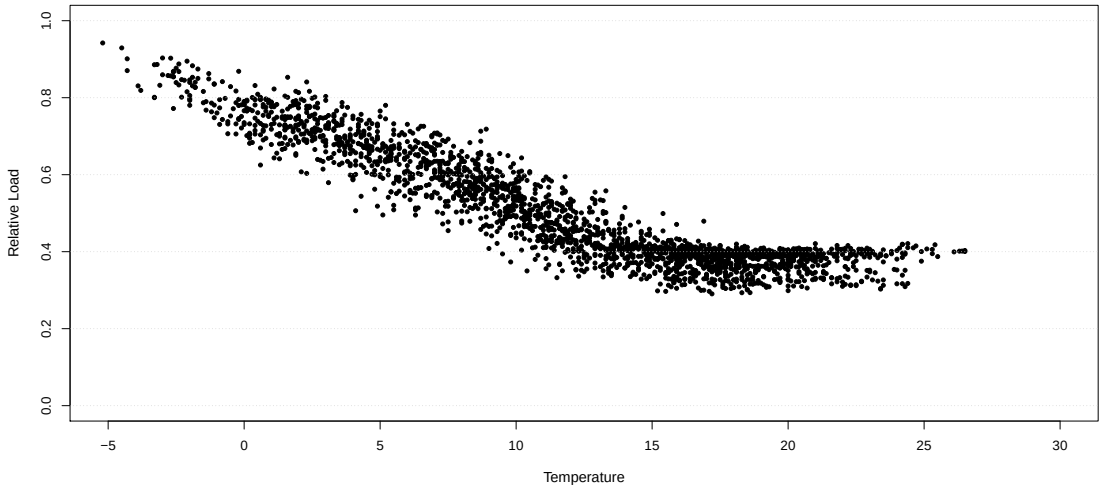

Figure 2: Relative electricity load at 10:00 over 5 years for a population against temperatures. Relative load is the ratio between the electricity load and its maximum value. The load seems to increase linearly with the temperature below a certain threshold.

The non-linearity of the model (see Seber and Wild, 2003, for a general presentation) comes both from the threshold u (which can be seen as a change-point in the model) and the shape parameters ψ_j. Gallant and Fuller (1973) and Gallant (1975) focused on least squares estimations in the case of non-linear models while Hinkley (1971) dealt with the special case of the two-phase regression (single change-point model). Bayesian linear regression was studied in Minka (1999) where piecewise linear regression was mentionned too. Marriott and Spencer (2001) worked on the predictive analysis from a Bayesian linear regression model when proper conjugate priors were used and derived expressions of the posterior predictive densities in such a case. Harrison and Stevens (1976) focused on Bayesian forecasting and exposed results for dynamic linear models from the Kalman filter point of view. This paper presents an application of a hierarchical Bayesian non-linear model (see Gelman and Hill, 2007, for a general review on the subject of huerarchical models) to the intraday electricity load forecasting problem.

Due to its highly periodic nature, the EVENTAIL model that we presented above typically requires 4 or 5 years of data to provide satisfactory estimations that will lead to reasonable predictions. In the situation where a study over a new population was started a year ago, we

would however only have a year –or possibly even less– of data by now. Fitting an appropriate model i.e. a model that would provide accurate predictions to such a reduced dataset is admittedly not an easy task : estimating the parameters of model via maximum likelihood will usually result in getting very high prediction errors (typical overfit situation).

Although electricity load curves may largely differ from one population to another, they may also share some common features. The latter case happens when the population studied is an aggregation of non-homogeneous subpopulations. This is the kind of information the user might be able to know, without either being able to formally express it or being able to take advantage of it. This is the reason why we are interested in building a prior distribution from a first population to help estimate the model over a second, exhibiting at least some similarities with the first and for which only a small dataset is available to the user.

The paper is organised as follows. In section 2 we focus on the general methodology and describe the way we carried our experimentations. In section 3 we present a non-informative model that will be applied for modelling the load when N is large (4 or 5 years). In section 4 we propose a hierarchical prior distribution to take into account information learnt on the load observed over a long period of time on a population having a similar behaviour. In both sections 3 and 4 we also provide ad hoc MCMC algorithms to estimate the mean and variance of the associated posterior distributions. These algorithms were needed for our tests on simulated datasets since we had to run them several times on many different simulated datasets and all-in-one solutions like WinBUGS (see Lunn et al., 2000, for documentation) would offer poor mixing Markov chains and require long execution times without any ability to finetune the algorithms used within. In sections 5 and 6 we use these algorithms to illustrate and validate our approach on simulated datasets and real datasets : we show the contribution of the prior over the precision of both the estimated parameters and the forecasts.

2 Methodology

2.1 The general model

In this subsection we present EVENTAIL as a particular case of a more general class of models. It can be written in the following condensed way which will be more convenient to work with : for $t = 1, \ldots, N$,

$$y_t = (A_t.\alpha) \times (B_t.\beta + C_t) + \gamma(T_t - u)\mathbb{1}_{[T_t, +\infty[}(u) + \epsilon_t \tag{2}$$

where $\epsilon_1, \ldots, \epsilon_N$ are independent and identically distributed with common distribution $\mathcal{N}(0, \sigma^2)$. The matrices A of size $N \times d_A$, B of size $N \times d_\beta$, C of size $N \times 1$, and T of size $N \times 1$ are known exogenous variables while the parameters of the model to be estimated are

$$\alpha \in \mathbb{R}^{d_\alpha}, \qquad \beta \in B_+^{d_\beta}(0,1), \qquad (\gamma, u) \in R^* \times [\underline{u}, \overline{u}] \qquad \sigma^2 \in \mathbb{R}_+^*$$

where

$$B_+^{d_\beta}(0,1) = \left\{\beta \in (\mathbb{R}_+)^{d_\beta}, \quad \|\beta\|_1 \leq 1\right\}$$

is the positive quadrant of the $\|\cdot\|_1$-unit ball of dimension d_β.

For the sake of completeness we give the explicit relationships between the general model and the EVENTAIL model. Let $d_\alpha = 2d_{11} + d_{12}$ and $d_\beta = d_2 - 1$. The matrices A of size $N \times d_\alpha$, B of size $N \times d_\beta$ and C of size $N \times 1$ for the EVENTAIL model are then defined as follows :

$$A = \begin{bmatrix} \cos\left(\frac{2\pi\times 1\times 1}{365.25}\right) & \cdots & \cos\left(\frac{2\pi\times d_{11}\times 1}{365.25}\right) & \sin\left(\frac{2\pi\times 1\times 1}{365.25}\right) & \cdots & \sin\left(\frac{2\pi\times d_{11}\times 1}{365.25}\right) & \mathbb{1}_{\Omega_1}(1) & \cdots & \mathbb{1}_{\Omega_{d_{12}}}(1) \\ \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ \cos\left(\frac{2\pi\times 1\times N}{365.25}\right) & \cdots & \cos\left(\frac{2\pi\times d_{11}\times N}{365.25}\right) & \sin\left(\frac{2\pi\times 1\times N}{365.25}\right) & \cdots & \sin\left(\frac{2\pi\times d_{11}\times N}{365.25}\right) & \mathbb{1}_{\Omega_1}(N) & \cdots & \mathbb{1}_{\Omega_{d_{12}}}(N) \end{bmatrix},$$

$$B = \begin{bmatrix} \mathbb{1}_{\Psi_1}(1) - \mathbb{1}_{\Psi_{d_2}}(1) & \cdots & \mathbb{1}_{\Psi_{d_2-1}}(1) - \mathbb{1}_{\Psi_{d_2}}(1) \\ \vdots & & \vdots \\ \mathbb{1}_{\Psi_1}(N) - \mathbb{1}_{\Psi_{d_2}}(N) & \cdots & \mathbb{1}_{\Psi_{d_2-1}}(N) - \mathbb{1}_{\Psi_{d_2}}(N) \end{bmatrix}, \qquad C = \begin{bmatrix} \mathbb{1}_{\Psi_{d_2}}(1) \\ \vdots \\ \mathbb{1}_{\Psi_{d_2}}(N) \end{bmatrix}.$$

The correspondence between the parameters of the model EVENTAIL and the general model is thus obvious : α regroups the parameters $z_j^{\cos}$, $z_j^{\sin}$, ω_j, β is a reparametrisation of ψ (that has a practical advantage over ψ, as we discuss hereafter) and (γ, u) corresponds to (g, u) without any change. Notice that the new parametrisation via β instead of ψ transforms the identifiability condition $\|\psi\|_1 = 1$ in the model (1) into the equivalent restriction $\|\beta\|_1 \leq 1$ in the model (2). We thus go from d_2 parameters linked together via a single linear equation in (1) down to $d_\beta = d_2 - 1$ parameters linearly independent in (2). It will be especially convenient in the next sections when we will need to build a prior for these parameters as we will be able to use a non-degenerated truncated Gaussian distribution.

2.2 Outline

Hereafter, we denote $\mathcal{B}$ a "short" dataset over which we would like to estimate the model and we denote $\mathcal{A}$ a "large" dataset known or thought to share some common features with $\mathcal{B}$. The Bayesian framework is an almost immediate choice for whoever wants to enjoy the benefits of some prior information, and since we aim at estimating a model over a dataset using information from another, it certainly imposed itself there.

Let us define here some notations that we shall keep throughout this paper. Denote first the regression data of day t as $\mathcal{D}_t = (A_{t\cdot}, B_{t\cdot}, C_t, T_t)$, and the regression data of days $1, \ldots, N$ as $\mathcal{D} = (\mathcal{D}_1, \ldots, \mathcal{D}_N)$. Denote then the parameters of the model

$$\theta = (\alpha, \beta, \gamma, u, \sigma^2) =: (\eta, \sigma^2) \in \mathbb{R}^d.$$

with $d = d_\alpha + d_\beta + 3$. Finally introduce $\mu(\eta|\mathcal{D})$ the expected values of the model for given data, a vector in $\mathbb{R}^N$ with coordinates

$$\mu_t(\eta|\mathcal{D}) = A_{t\cdot}\alpha \times (B_{t\cdot}\beta + C_t) + \gamma(T_t - u)\mathbb{1}_{[T_t, +\infty[}(u), \quad t = 1, \ldots, N.$$

The likelihood of the model is given by

$$L(y|\theta, \mathcal{D}) \propto \sigma^{-N} \exp\left(-\frac{1}{2}\tau\|y - \mu(\eta|\mathcal{D})\|_2^2\right) \mathbb{1}_{[0,1]\times[\underline{u}, \overline{u}]\times\mathbb{R}_+^*}(\|\beta\|_1, u, \sigma^2).$$

We provide a two–stage method to help improve parameter estimations and model predictions over $\mathcal{B}$ with the help of $\mathcal{A}$, which goes as follows :

Stage 1. we build a non-informative prior over θ, $\pi^{\mathcal{A}}(\theta)$ and use it for the estimation of the parameters of the Bayesian model over $\mathcal{A}$ whose observations we denote $y^{\mathcal{A}} = (y_1^{\mathcal{A}}, \ldots, y_{N_{\mathcal{A}}}^{\mathcal{A}})$. We then calculate the corresponding posterior distribution given by

$$\pi^{\mathcal{A}}(\theta|y_1^{\mathcal{A}}, \ldots, y_{N_{\mathcal{A}}}^{\mathcal{A}}, \mathcal{D}^{\mathcal{A}}) \propto L(y^{\mathcal{A}}|\theta, \mathcal{D}^{\mathcal{A}})\pi^{\mathcal{A}}(\theta),$$

and denote its posterior mean and variance

$$\mu^{\mathcal{A}} = \mathbb{E}^{\pi^{\mathcal{A}}}[\eta|y_1^{\mathcal{A}}, \ldots, y_{N_{\mathcal{A}}}^{\mathcal{A}}],$$
$$\Sigma^{\mathcal{A}} = \mathrm{Var}^{\pi^{\mathcal{A}}}[\eta|y_1^{\mathcal{A}}, \ldots, y_{N_{\mathcal{A}}}^{\mathcal{A}}].$$

Stage 2. we build a hierarchical prior over η (all the parameters of the model except σ^2) based on $\mu^{\mathcal{A}}$ and $\Sigma^{\mathcal{A}}$ to help achieve better estimations of the parameters on the dataset $\mathcal{B}$. We chose to use a hierarchical prior over η of the following form

$$\eta|k, l \sim \mathcal{N}(M^{\mathcal{A}}k, l^{-1}\Sigma^{\mathcal{A}})$$
$$k|q, r \sim \mathcal{N}(q(1, \dots, 1)', r^{-1}I_d)$$

where l, q and r are themselves hyperparameters (just like k is) for which we decide to use non-informative (vague) prior distributions and where $M^{\mathcal{A}} = \text{diag}(\mu^{\mathcal{A}})$ is the diagonal matrix whose diagonal coefficients are given by $\mu^{\mathcal{A}}$. Finally we estimate the associated posterior distribution and the corresponding posterior mean and variance. The coordinates of k can be seen as a similarity coefficients since it is used to scale the mean learnt on $\mathcal{A}$ to fit on $\mathcal{B}$, while q can be interpreted as a global similarity measure between the two datasets. The prior mean of q is of course forced to $\mathbb{E}[q] = 1$ to reflect the prior knowledge that the datasets are somehow similar, and the variance of the prior distribution of q could be reduced, depending on the confidence we have over the similarity between the datasets. However we chose not to, to keep the procedure we describe from requiring any delicate or subjective adjustments. Note that even though the priors from the upper layers of the model are vague, the correlations between parameters learnt on the dataset $\mathcal{A}$ remain untouched and are thus directly used in this hierarchical prior.

3 Non-informative approach

In this section we present the non-informative framework that can be applied to any dataset, long or short. We first show that the use of a non-informative prior distribution leads to a proper posterior distribution (see Proposition 1). We then propose an MCMC algorithm to (approximately) simulate from this posterior distribution and be able to retrieve posterior estimations of the different parameters.

3.1 Prior and posterior distributions

We use the following non-informative prior

$$\pi(\theta) \propto \sigma^{-2}.$$

This prior is non-informative in the sense that it matches Jeffreys' prior distribution on σ^2 for a Gaussian linear regression and matches Laplace's flat prior on the other parameters. It leads to the following posterior distribution

$$\begin{aligned}\pi(\theta|y, \mathcal{D}) &\propto L(y|\theta, \mathcal{D})\pi(\theta) \\ &\propto \sigma^{-N-2} \exp\left(-\frac{1}{2}\sigma^{-2}\|y - \mu(\eta|\mathcal{D})\|_2^2\right) \mathbb{1}_{[0,\,1]\times[\underline{u},\,\overline{u}]\times\mathbb{R}_+^*}(\|\beta\|_1, u, \sigma^2). \end{aligned} \tag{3}$$

Proposition 1. *For $(\beta, u) \in B_+^{d_\beta}(0,1) \times [\underline{u},\, \overline{u}]$ denote $A_*(\beta, u)$ the matrix whose rows are*

$$(A_*)_{t\cdot}(\beta, u) = \left[(B_{t\cdot}\beta + C_t)A_{t\cdot}, (T_t - u)\mathbb{1}_{[T_t,\,+\infty[}(u)\right], \quad t = 1, \dots, N,$$

and suppose $A_*'(b, u)A_*(b, u)$ has full rank for every $(\beta, u) \in B_+^{d_\beta}(0,1)\times[\underline{u},\, \overline{u}]$. Assume furthermore that $N > d_\alpha + 1$ and that $(y_1, \dots, y_N)$ are observations coming from the model (2) and the posterior measure (3) is then a well-defined (proper) probability distribution.

Proof. Notice first that

$$\int \pi(\eta, \sigma^2|y, \mathcal{D})\, \mathrm{d}\sigma^2 \propto \|y - \mu(\eta|\mathcal{D})\|_2^{-N} \mathbb{1}_{[0,\,1]}(\|b\|_1) \mathbb{1}_{[\underline{u},\,\overline{u}]}(u) \quad \text{for almost every } y,$$

and observe then that

$$\|y - \mu(\eta|\mathcal{D})\|_2^2 = \sum_{t=1}^{N} \left[y_t - (B_{t\cdot}\beta + C_t) A_{t\cdot}\alpha - (T_t - u)\mathbb{1}_{[T_t,\,+\infty[}(u)\gamma \right]^2.$$

Let $(\beta_0, u_0) \in B_+^{d_\beta}(0,1) \times [\underline{u},\, \overline{u}]$ and denote $\alpha_* = (\alpha, \gamma)$. We write

$$\begin{aligned} \|y - \mu((\alpha, \beta_0, \gamma, u_0)|\mathcal{D})\|_2^2 &= \sum_{t=1}^{N} \left[y_t - (B_{t\cdot}\beta_0 + C_t) A_{t\cdot}\alpha - (T_t - u_0)\mathbb{1}_{[T_t,\,+\infty[}(u_0)\gamma \right]^2 \\ &= \|y - A_*(\beta_0, u_0)\alpha_*\|_2^2, \end{aligned}$$

and thus obtain the following equivalence, as $(\beta, u) \to (\beta_0, u_0)$ and $\|\alpha_*\|_2 \to +\infty$

$$\|y - \mu(\eta|\mathcal{D})\|_2^{-N} \sim \|y - A_*(\beta_0, u_0)\alpha_*\|_2^{-N}. \tag{4}$$

The triangular inequality applied to the right hand side of (4) gives

$$\|y - A_*(\beta_0, u_0)\alpha_*\|_2^{-N} \leq \big| \|y\|_2 - \|A_*(\beta_0, u_0)\alpha_*\|_2 \big|^{-N}. \tag{5}$$

Since $A_*'(\beta_0, u_0)A_*(\beta_0, u_0)$ has full rank, by straightforward algebra we get

$$\lambda \|\alpha_*\|_2^2 \leq \|A_*(\beta_0, u_0)\alpha_*\|_2^2, \tag{6}$$

where λ is the smallest eigenvalue $(A_*(\beta_0, u_0))' A_*(\beta_0, u_0)$ and is strictly positive. We can hence find an equivalent of the right hand side of (5) as $\|\alpha_*\|_2 \to +\infty$, which is

$$\big| \|y\|_2 - \|A_*(\beta_0, u_0)\alpha_*\|_2 \big|^{-N} \sim \lambda^{-N/2} \|\alpha_*\|_2^{-N}. \tag{7}$$

Combining (4), (5) and (7) together, we see that the integrability of the left hand side of (4) as $(\beta, u) \to (\beta_0, u_0)$ and $\|\alpha_*\|_2 \to +\infty$ is directly implied by that of $\|\alpha_*\|_2^{-N}$. The latter is of course immediate for $N > d_\alpha + 1$ as can be seen via a quick cartesian to hyperspherical re-parametrisation.

The previous paragraph thus ensures the integrability of $\|y - \mu(\eta|\mathcal{D})\|_2^{-N}$ over sets of the form

$$\{(\beta, u) \in V((\beta_0, u_0)), \|\alpha_*\|_2 \in]M(\beta_0, u_0),\, +\infty[\}, \quad \forall (\beta_0, u_0) \in B_+^{d_\beta}(0,1) \times [\underline{u},\, \overline{u}]$$

where the subset $V((b_0, u_0))$ is an open neighbourhood of (β_0, u_0) and $M(\beta_0, u_0)$ is a real number depending on (β_0, u_0). By compacity of $B_+^{d_\beta}(0,1) \times [\underline{u},\, \overline{u}]$ there exists a finite union of such $V((\beta_i, u_i))$ that covers $B_+^{d_\beta}(0,1) \times [\underline{u},\, \overline{u}]$. Denoting M the maximum of $M(\beta_i, u_i)$ over the corresponding finite subset of (β_i, u_i), we finally obtain the integrability of $\|y - \mu(\eta|\mathcal{D})\|_2^{-N}$ over $\{(\beta, u) \in B_+^{d_\beta}(0,1), \|\alpha_*\| \in]M,\, +\infty[\}$.

The integrability of $\|y - \mu(\eta|\mathcal{D})\|_2^{-N}$ over $\{(\beta, u) \in B_+^{d_\beta}(0,1), \|\alpha_*\| \in [0,\, M]\}$ is trivial, recalling that $\eta \mapsto \|y - \mu(\eta|\mathcal{D})\|_2$ is continuous and does not vanish over this compact for almost every y, meaning its inverse enjoys the same properties. □

Remark 2. *The condition "$A_*'A_*$ has full rank" mentionned above is typically verified in our applications for the regressors used in the* EVENTAIL *model. To see this, call "vector of heating degrees" the vector whose coordinates are $(T_t - u)\mathbb{1}_{[T_t,\,+\infty[}(u)$, then not verifying the aforementionned condition is equivalent to saying that "there exists an index i and a threshold u such that the family of vectors formed by the regressors A and the vector of heating degrees is linearly dependant over the subset Ψ_i of the calendar".*

3.2 MCMC algorithm

We start this subsection by giving the different steps of the MCMC algorithm we used to simulate $(\theta_1, \ldots, \theta_M)$ according to the posterior distribution $\pi(\theta|y, \mathcal{D})$. This MCMC algorithm was developed because direct simulations from the posterior distribution were not possible. The justifications are given after the algorithm itself. Notice that the full conditional distributions of all the parameters but the threshold u appear to be common distributions. We thus used a Metropolis-within-Gibbs algorithm (see Marin and Robert, 2007, page 96, for a quick description) based on Gibbs sampling steps for every parameter but u for which we use a Metropolis-Hasting step based on a gaussian random walk proposal. The algorithm goes as follows :

Step 1. Initialise θ_1 such that $\pi(\theta_1|y, \mathcal{D}) \neq 0$

Step 2. For $t = 1, \ldots, M-1$, repeat

(i) Simulate σ^2_{t+1} conditionally to $(\alpha_t, \beta_t, \gamma_t, u_t, y, \mathcal{D})$ i.e.

$$\sigma^2_{t+1} \sim \mathcal{IG}\left(\frac{N}{2}, \frac{1}{2}\|y - \mu(\eta|\mathcal{D})\|_2^2\right)^{2}$$

(ii) Simulate γ_{t+1} conditionally to $(\alpha_t, \beta_t, u_t, \sigma^2_{t+1}, y, \mathcal{D})$ i.e.

$$\gamma_{t+1} \sim \mathcal{N}\left(\mu^{\gamma}_{t+1}, \Sigma^{\gamma}_{t+1}\right)$$

(iii) Simulate b_{t+1} conditionally to $(\alpha_t, \gamma_{t+1}, u_t, \sigma^2_{t+1}, y, \mathcal{D})$ i.e.

$$\beta_{t+1} \sim \mathcal{N}\left(\mu^{\beta}_{t+1}, \Sigma^{\beta}_{t+1}, B_+^{d_\beta}(0,1)\right)^{3}$$

(iv) Simulate a_{t+1} conditionally to $(\beta_{t+1}, \gamma_{t+1}, u_t, \sigma^2_{t+1}, y, \mathcal{D})$ i.e.

$$\alpha_{t+1} \sim \mathcal{N}\left(\mu^{\alpha}_{t+1}, \Sigma^{\alpha}_{t+1}\right)$$

(v) Simulate $\delta_t \sim \mathcal{N}(0, \Sigma_{\text{MH}})$, simulate $\upsilon_t \sim \mathcal{U}[0,\, 1]^{4}$ and define $\widetilde{u}_t = u_t + \delta_t$

- define $u_{t+1} = \widetilde{u}_t$ if

$$\upsilon_t < \frac{\pi(\widetilde{u}_t|\alpha_{t+1}, \beta_{t+1}, \gamma_{t+1}, \sigma^2_{t+1}, y, \mathcal{D})}{\pi(u_t|\alpha_{t+1}, \beta_{t+1}, \gamma_{t+1}, \sigma^2_{t+1}, y, \mathcal{D})}$$

- or $u_{t+1} = u_t$ otherwise

where the covariance matrix Σ_{MH} used in this last Metropolis-Hastings step is first estimated over a burn-in phase (the iterations coming from this phase are discarded), and then fixed to its estimated value "asymptotically optimally rescaled" for the final run by a factor $\left(\frac{2.38}{d}\right)^2$ (as recommended for Gaussian proposals in section 2 of Roberts and Rosenthal, 2001).

The justifications for each full conditional distribution used in the Gibbs sampling steps, including the explicit expressions of $\mu^{\alpha}_{t+1}, \Sigma^{\alpha}_{t+1}, \mu^{\beta}_{t+1}, \Sigma^{\beta}_{t+1}, \mu^{\gamma}_{t+1}$, and Σ^{γ}_{t+1}, are given in the following paragraphs. Lemma 3 below is a key element to these justifications.

---

[2] $\mathcal{IG}(a, b)$ is the inverse-gamma distribution with density $\frac{b^a}{\Gamma(a)}\frac{e^{-b/x}}{x^{a+1}}\mathbb{1}_{[0,\,+\infty[}(x)$.

[3] $\mathcal{N}(m, \Sigma, S)$ is the truncated Gaussian distribution on S with density proportional to $\exp\left(-\frac{1}{2}(x-m)'\Sigma^{-1}(x-m)\right)\mathbb{1}_S(x)$. Detailed explanations on how to simulate random variables with such a distribution are available in Robert (1995)

[4] $\mathcal{U}[a,\, b]$ is the uniform distribution with density $\frac{1}{b-a}\mathbb{1}_{[a,\,b]}(x)$.

Lemma 3. *Let X and Y be two random vectors respectively in $\mathbb{R}^d$ et $\mathbb{R}^n$ such as the conditional distribution of Y with regard to X is Gaussian*

$$Y|X \sim \mathcal{N}\left(Z + MX, \sigma^2 I_n\right)$$

with M matrix of size $n \times d$ that has full rank $d < n$, and let Z be a fixed vector in $\mathbb{R}^n$. The conditional distribution of X with regard to Y is then Gaussian too

$$X|Y \sim \mathcal{N}\left([M'M]^{-1}M'(Y-Z), \sigma^2 M'M\right).$$

Proof. Denoting $W = Y - Z$, straightforward algebra leads immediately to

$$\begin{aligned}(W-MX)'\sigma^2 I_n (W-MX) = \big((M'M)^{-1}M'W - X\big)'\sigma^2(M'M)^{-1}\big((M'M)^{-1}M'W - X\big) \\ + \Big(W'\sigma^2 I_n W - \big((M'M)^{-1}M'W\big)'\sigma^2(M'M)^{-1}\big((M'M)^{-1}M'W\big)\Big)\end{aligned}$$

where the second term on the right hand side of the equation does not depend on X. □

Full conditional distribution of α. The full conditional distribution of α can directly be deduced from both the prior and the likelihood contributions to it. Denote n the size of the vector α, and $\theta\backslash\alpha$ the vector θ from which the coordinates corresponding to α have been removed.

Let us first observe that, since the prior distribution we are using on α is flat, the full conditional distribution of α is in fact proportional to the likelihood function (seen as a function of α). Now considering the likelihood contribution, we write

$$\pi(\alpha|\eta\backslash\alpha, y, \mathcal{D}) \propto \exp\left(-\frac{1}{2}\sigma^{-2}\|y - \mu(\eta|\mathcal{D})\|_2^2\right)$$

Let now L_α be the diagonal matrix whose diagonal coefficients are given by

$$(L_\alpha)_{tt} = B_{t\cdot}\beta + C_t, t = 1, \ldots, N,$$

let Z_α be the vector whose coordinates are given by

$$(Z_\alpha)_t = \gamma(T_t - u)\mathbb{1}_{[T_t, +\infty[}(u), \quad t = 1, \ldots, N,$$

and denote M_α the matrix $M_\alpha = L_\alpha A$. We can now rewrite μ and get

$$\pi(\alpha|\theta\backslash\alpha, y, \mathcal{D}) \propto \exp\left(-\frac{1}{2}\sigma^{-2}\|y - (Z_\alpha + M_\alpha\alpha)\|_2^2\right).$$

Using Lemma 3, it is then straightforward to see that the full conditional distribution of α is Gaussian

$$\alpha|\theta\backslash\alpha, y, \mathcal{D} \sim \mathcal{N}\left(\mu^\alpha, \Sigma^\alpha\right) \tag{8}$$

where

$$\begin{pmatrix} \mu^\alpha \\ \Sigma^\alpha \end{pmatrix} = \begin{pmatrix} [M'_\alpha M_\alpha]^{-1}M'_\alpha(y - Z_\alpha) \\ \sigma^2 M'_\alpha M_\alpha \end{pmatrix}.$$

Full conditional distribution of β. Using similar arguments, we obtain the full conditional distribution of β. Namely, denoting Z_β the vector whose coordinates are given by

$$(Z_\beta)_t = (A\alpha)_t C_t + \gamma(T_t - u)\mathbb{1}_{[T_t, +\infty[}(u),$$

and calling $M_\beta = L_\beta B$ where L_β is the diagonal matrix whose diagonal is $A\alpha$, we obtain the truncated Gaussian distribution

$$\beta|\theta\backslash\beta, y, \mathcal{D} \sim \mathcal{N}(\mu^\beta, \Sigma^\beta, B_+^{d_\beta}(0,1)) \tag{9}$$

where

$$\begin{pmatrix} \mu^\beta \\ \Sigma^\beta \end{pmatrix} = \begin{pmatrix} [M'_\beta M_\beta]^{-1} M'_\beta (y - Z_\beta) \\ \sigma^2 M'_\beta M_\beta \end{pmatrix}.$$

Full conditional distribution of γ. Using once again similar arguments, we obtain the full conditional distribution of γ. Namely, denoting Z_γ the vector whose coordinates are given by

$$(Z_\gamma)_t = (A\alpha)_t((B\beta)_t + C_t),$$

and calling M_γ the vector whose coordinates are $(T_t - u)\mathbb{1}_{[T_t, +\infty[}(u)$ we obtain the Gaussian distribution

$$\gamma|\theta\backslash\gamma, y, \mathcal{D} \sim \mathcal{N}(\mu^\gamma, \Sigma^\gamma) \tag{10}$$

where

$$\begin{pmatrix} \mu^\gamma \\ \Sigma^\gamma \end{pmatrix} = \begin{pmatrix} [M'_\gamma M_\gamma]^{-1} M'_\gamma (y - Z_\gamma) \\ \sigma^2 M'_\gamma M_\gamma \end{pmatrix}.$$

Full conditional distribution of σ^2. No calculations are required, as we immediately identify an inverse-gamma distribution from (3).

4 Hierarchical approach

In this section we present the hierarchical prior we build from $\mathcal{A}$ to improve our results on $\mathcal{B}$. To be able to build it, we first applied the non-informative approach to the large dataset $\mathcal{A}$ and thus collected $\mu^\mathcal{A}$ and $\Sigma^\mathcal{A}$ the posterior mean and posterior variance of η on $\mathcal{A}$. Hereafter we denote $M^\mathcal{A} = \text{diag}(\mu^\mathcal{A})$. The hierarchical prior that we propose in the next subsection introduces new parameters to model the similarity between the two datasets. Note that for the sake of clarity, we drop the $^\mathcal{B}$ notation : when not explicitly specified, the dataset, data and observations as well as the prior and posterior distributions we refer to in this section will be those corresponding to $\mathcal{B}$. As in the previous section, we first describe the hierarchical prior we use, and prove that it leads to a proper posterior distribution (see Proposition 4). We then propose an MCMC algorithm to (approximately) simulate from this posterior distribution and be able to retrieve posterior estimations of the different parameters as we did when we worked with the non-informative prior.

4.1 Prior and posterior distributions

Instead of using the obvious and far from robust prior

$$\eta \sim \mathcal{N}(\mu^\mathcal{A}, \Sigma^\mathcal{A})$$

we introduce hyperparameters $(k, l) \in \mathbb{R}^d \times \mathbb{R}$ and $(q, r) \in \mathbb{R} \times \mathbb{R}_+^*$ such that

$$\begin{aligned} \eta|k, l &\sim \mathcal{N}(M^\mathcal{A} k, l^{-1}\Sigma^\mathcal{A}) \\ k|q, r &\sim \mathcal{N}(q(1, \dots, 1)', r^{-1} I_d) \end{aligned}$$

to allow for more robustness. The coordinates of the vector k can be interpreted as similarity coefficients between parameters of $\mathcal{A}$ and $\mathcal{B}$ and the strictly positive scalar l can be seen as a way to alternatively weaken or strengthen the covariance matrix as needed. q and r are more general indicators of how close $\mathcal{A}$ and $\mathcal{B}$ are, q corresponding to the mean of the coordinates of k and r being their inverse-variance. l, q, r and σ^2 of course require a prior distribution too. For σ^2 we

use the same non-informative prior that we used when considering the non-informative approach (i.e. $\pi(\sigma^2) = \sigma^{-2}$). For the three other parameters we use :

$$\begin{aligned} l &\sim \mathcal{G}(a_l, b_l) \\ q &\sim \mathcal{N}(1, \sigma_q^2) \\ r &\sim \mathcal{G}(a_r, b_r) \end{aligned}$$

where a_l, b_l, a_r, b_r and σ_q^2 are fixed positive real numbers such that the prior distribution on l, q and r are vague. These prior distributions are chosen because of their conjugacy properties (as will be seen in the MCMC algorithm). The vagueness requirement that we impose on these priors is motivated by the fact that we want to keep as general a framework as possible without having to tweak each and every prior coefficient for different applications.

The hierarchical prior that we use is built as follows :

$$\pi(\theta, k, l, q, r) \propto \pi(\eta|k, l)\pi(k|q, r)\pi(l)\pi(q)\pi(r)\pi(\sigma^2) \tag{11}$$

with

$$\begin{aligned} \pi(\sigma^2) &\propto \sigma^{-2} \\ \pi(\eta|k, l) &\propto l^{\frac{d}{2}} \exp\left(-\frac{1}{2}(\theta - M^{\mathcal{A}}k)' l (\Sigma^{\mathcal{A}})^{-1}(\theta - M^{\mathcal{A}}k)\right) \\ \pi(k|q, r) &\propto |r|^{\frac{d}{2}} \exp\left(-\frac{1}{2} r \sum_{i=1}^{d} (k_i - q)^2\right) \\ \pi(l) &\propto l^{a_l - 1} \exp(-b_l l)\, \mathbb{1}_{\mathbb{R}_+^*}(l) \\ \pi(q) &\propto |\sigma_q^{-2}|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\sigma_q^{-2}(q-1)^2\right) \\ \pi(r) &\propto r^{a_r - 1} \exp(-b_r r)\, \mathbb{1}_{\mathbb{R}_+^*}(r). \end{aligned}$$

The posterior measure is hence given by

$$\begin{aligned} \pi(\theta, k, l, q, r|y, \mathcal{D}) &\propto f(y|\theta, \mathcal{D})\pi(\theta, k, l, q, r) \\ &\propto \sigma^{-N-2} \exp\left(-\frac{1}{2}\sigma^{-2}\|y - \mu(\eta|\mathcal{D})\|_2^2\right) \mathbb{1}_{[0,\,1]\times[\underline{u},\,\overline{u}]\times\mathbb{R}_+^*}(\|\beta\|_1, u, \sigma^2) \\ &\quad \times |r|^{\frac{d}{2}} \exp\left(-\frac{1}{2} r \sum_{i=1}^{d} (k_i - q)^2\right) l^{a_l - 1} \exp(-b_l l)\, \mathbb{1}_{\mathbb{R}_+^*}(l) \\ &\quad \times |\sigma_q^{-2}|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\sigma_q^{-2}(q-1)^2\right) r^{a_r - 1} \exp(-b_r r)\, \mathbb{1}_{\mathbb{R}_+^*}(r). \end{aligned} \tag{12}$$

Proposition 4. *For $(\beta, u) \in B_+^{d_\beta}(0,1) \times [\underline{u},\, \overline{u}]$ denote $A_*(\beta, u)$ the matrix whose rows are*

$$(A_*)_{t\cdot}(\beta, u) = \left[(B_{t\cdot}\beta + C_t)A_{t\cdot}, (T_t - u)\mathbb{1}_{[T_t,\,+\infty[}(u)\right], \quad t = 1, \ldots, N,$$

and suppose $A_*'(b, u)A_*(b, u)$ has full rank for every $(\beta, u) \in B_+^{d_\beta}(0,1)\times[\underline{u},\, \overline{u}]$. Assume furthermore that $N > d_\alpha + 1$ and that $(y_1, \ldots, y_N)$ are observations coming from the model (2) and the posterior measure (12) is then a well-defined (proper) probability distribution.

Proof. First notice that

$$\int \pi(\theta, k, l, q, r|y, \mathcal{D})\, \mathrm{d}\sigma^2 \propto \|y - \mu(\eta|\mathcal{D})\|_2^{-N} \mathbb{1}_{[0,\,1]}(\|\beta\|_1)\mathbb{1}_{[\underline{u},\,\overline{u}]}(u) \times \pi(\eta|k, l)\pi(k|q, r)\pi(l)\pi(q)\pi(r),$$

for almost every y and that the function $\theta \mapsto \|y - \mu(\eta|\mathcal{D})\|_2^{-N}$ is bounded, for almost every y. The posterior integrability is therefore trivial as long as $\pi(\eta|k, l)\pi(k|q, r)\pi(l)\pi(q)\pi(r)$ itself is a proper distribution which is the case here. □

4.2 MCMC Algorithm

We start this subsection by giving the different steps of the MCMC algorithm we used to simulate $(\theta_1, \ldots, \theta_M)$ according to the posterior distribution $\pi(\theta|y, \mathcal{D})$. The justifications are given after the algorithm itself. The algorithm goes as follow :

`Step 1. Initialise` θ_1 `such that` $\pi(\theta_1|y, \mathcal{D}) \neq 0$

`Step 2. For` $t = 1, \ldots, M-1$`, repeat`

`(i) Simulate` σ^2_{t+1} `conditionally to` $(\alpha_t, \beta_t, \gamma_t, u_t, k_t, l_t, q_t, r_t, y, \mathcal{D})$ `i.e.`

$$\sigma^2_{t+1} \sim \mathcal{IG}\left(\frac{N}{2}, \frac{1}{2}\|y - \mu(\eta|\mathcal{D})\|_2^2\right)$$

`(ii) Simulate` r_{t+1} `conditionally to` $(\alpha_t, \beta_t, \gamma_t, u_t, \sigma^2_{t+1}, k_t, l_t, q_t, y, \mathcal{D})$ `i.e.`

$$r_{t+1} \sim \mathcal{G}\left(a_r + \frac{d}{2}, b_r + \frac{1}{2}\sum_{i=1}^{d}(k_i - q)^2\right)$$ [5]

`(iii) Simulate` q_{t+1} `conditionally to` $(\alpha_t, \beta_t, \gamma_t, u_t, \sigma^2_{t+1}, k_t, l_t, r_{t+1}, y, \mathcal{D})$ `i.e.`

$$q_{t+1} \sim \mathcal{N}\left([\sigma_q^{-2} + rd]^{-1}(\sigma_q^{-2} \times 1 + r\sum_{i=1}^{d} k_i), [\sigma_q^{-2} + rd]^{-1}\right)$$

`(iv) Simulate` l_{t+1} `conditionally to` $(\alpha_t, \beta_t, \gamma_t, u_t, \sigma^2_{t+1}, k_t, q_{t+1}, r_{t+1}, y, \mathcal{D})$ `i.e.`

$$l_{t+1} \sim \mathcal{G}\left(a_l + \frac{d}{2}, b_l + \frac{1}{2}(\eta_t - M^{\mathcal{A}}k_t)'(\Sigma^{\mathcal{A}})^{-1}(\eta_t - M^{\mathcal{A}}k_t)\right)$$

`(v) Simulate` k_{t+1} `conditionally to` $(\alpha_t, \beta_t, \gamma_t, u_t, \sigma^2_{t+1}, l_{t+1}, q_{t+1}, r_{t+1}, y, \mathcal{D})$ `i.e.`

$$k_{t+1} \sim \mathcal{N}\left(\mu^k_{t+1}, \Sigma^k_{t+1}\right)$$

`(vi) Simulate` γ_{t+1} `conditionally to` $(\alpha_t, \beta_t, u_t, \sigma^2_{t+1}, k_{t+1}, l_{t+1}, q_{t+1}, r_{t+1}, y, \mathcal{D})$ `i.e.`

$$\gamma_{t+1} \sim \mathcal{N}\left(\mu^g_{t+1}, \Sigma^g_{t+1}\right)$$

`(vii) Simulate` β_{t+1} `conditionally to` $(\alpha_t, \gamma_{t+1}, u_t, \sigma^2_{t+1}, k_{t+1}, l_{t+1}, q_{t+1}, r_{t+1}, y, \mathcal{D})$ `i.e.`

$$\beta_{t+1} \sim \mathcal{N}\left(\mu^b_{t+1}, \Sigma^b_{t+1}, B_+^{d_\beta}(0, 1)\right)$$

`(viii) Simulate` α_{t+1} `conditionally to` $(\beta_{t+1}, \gamma_{t+1}, u_t, \sigma^2_{t+1}, k_{t+1}, l_{t+1}, q_{t+1}, r_{t+1}, y, \mathcal{D})$ `i.e.`

$$\alpha_{t+1} \sim \mathcal{N}\left(\mu^a_{t+1}, \Sigma^a_{t+1}\right)$$

`(ix) Simulate` $\delta_t \sim \mathcal{N}(0, \Sigma_{\text{MH}})$, $\quad \upsilon_t \sim \mathcal{U}[0, 1]$ `and define` $\widetilde{u}_t = u_t + \delta_t$

- `define` $u_{t+1} = \widetilde{u}_t$ `if`

$$\upsilon_t < \frac{\pi(\widetilde{u}_t|\alpha_{t+1}, \beta_{t+1}, \gamma_{t+1}, \sigma^2_{t+1}, k_{t+1}, l_{t+1}, q_{t+1}, r_{t+1}, y, \mathcal{D})}{\pi(u_t|\alpha_{t+1}, \beta_{t+1}, \gamma_{t+1}, \sigma^2_{t+1}, k_{t+1}, l_{t+1}, q_{t+1}, r_{t+1}, y, \mathcal{D})}$$

- `or` $u_{t+1} = u_t$ `otherwise`

[5] $\mathcal{G}(a, b)$ is the gamma distribution with density $\frac{b^a}{\Gamma(a)}\frac{e^{-bx}}{x^{a-1}}\mathbb{1}_{[0,\,+\infty[}(x)$.

where the covariance matrix Σ_{MH} used in the Metropolis-Hastings step is first estimated over a burn-in phase, and then fixed to its rescaled estimated value for the real run as in the non-informative approach.

The justifications for each full conditional distribution used in the Gibbs sampling steps, including the explicit expressions of $\mu^{\alpha}_{t+1}, \Sigma^{\alpha}_{t+1}, \mu^{\beta}_{t+1}, \Sigma^{\beta}_{t+1}, \mu^{\gamma}_{t+1}, \Sigma^{\gamma}_{t+1}, \mu^{k}_{t+1}$ and Σ^{k}_{t+1}, are given in the following paragraphs. To obtain these full conditional distributions, we will make use of Lemma 3 again, as well as Lemmas 6 and 7, both given below.

Definition 5 (Gaussian conjugacy operator)**.** *We define the (commutative and associative) operator $*$ as*

$$\begin{pmatrix} \mu_1 \\ \Sigma_1 \end{pmatrix} * \begin{pmatrix} \mu_2 \\ \Sigma_2 \end{pmatrix} = \begin{pmatrix} [\Sigma_1^{-1} + \Sigma_2^{-1}]^{-1}(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2) \\ [\Sigma_1^{-1} + \Sigma_2^{-1}]^{-1} \end{pmatrix}$$

for any vectors μ_1 and μ_2 in $\mathbb{R}^d$, for any symmetric positive definite matrices Σ_1 and Σ_2 of size $d \times d$.

Lemma 6 (Conjugacy)**.** *Let X_1 and X_2 be two random truncated Gaussian vectors in $\mathbb{R}^d$*

$$X_1 \sim \mathcal{N}(\mu_1, \Sigma_1, S_1)$$
$$X_2 \sim \mathcal{N}(\mu_2, \Sigma_2, S_2)$$

and denote f_1 and f_2 their respective densities, then $f_1 f_2$ is integrable. Let furthermore Y be a random variable with density $g(y) \propto f_1(y) f_2(y)$, then Y has truncated Gaussian distribution

$$Y \sim \mathcal{N}(\mu, \Sigma, S_1 \cap S_2)$$

where

$$\begin{pmatrix} \mu \\ \Sigma \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \Sigma_1 \end{pmatrix} * \begin{pmatrix} \mu_2 \\ \Sigma_2 \end{pmatrix}$$

and this result easily extends to any finite number of random truncated (or not) Gaussian vectors.

Lemma 7 (Conditional distribution)**.** *Let X be a random Gaussian vector in $\mathbb{R}^d$*

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} R & S \\ S' & T \end{bmatrix}^{-1} \right)$$

and X_1 and X_2 the projections of X over its d_1 first and d_2 last coordinates ($d = d_1 + d_2$). The conditional distribution of X_1 with regard to X_2 is then Gaussian

$$X_1 | X_2 \sim \mathcal{N}(\mu_1 - R^{-1}S(X_2 - \mu_2), R^{-1})$$

Full conditional distribution of α. The full conditional distribution of α can directly be deduced from both the prior and the likelihood contributions to it. Denote $\theta_* = (\theta, k, l, q, r)$, and write the full conditional distribution of α as

$$\pi(\alpha | \theta_* \backslash \alpha, y, \mathcal{D}) \propto g_L(\alpha) \cdot g_p(\alpha)$$

where $g_L(\alpha)$ is the contribution of the likelihood (seen as a function of α to the full conditional distribution) and $g_p(\alpha)$ is the contribution of the prior (seen as a function of α). We prove g_L and g_p both correspond to Gaussian distributions before using Lemma 6 to combine them into yet another Gaussian distribution.

1. Let us first consider the prior contribution g_p. Recall first that α only appears in the following component of the prior

$$\pi(\theta | k, l) \propto l^{\frac{d}{2}} \exp\left(-\frac{1}{2}(\theta - M^{\mathcal{A}}k)' l (\Sigma^{\mathcal{A}})^{-1} (\theta - M^{\mathcal{A}}k) \right),$$

which directly implies that

$$g_p(\alpha) \propto \exp\left(-\frac{1}{2}(\theta - M^{\mathcal{A}}k)'l(\Sigma^{\mathcal{A}})^{-1}(\theta - M^{\mathcal{A}}k)\right).$$

Denote $\mu = M^{\mathcal{A}}k$, $\Sigma = l^{-1}\Sigma^{\mathcal{A}}$ and denote μ_α and $\mu_{\eta\backslash\alpha}$ the vectors resulting from the extractions of the coordinates corresponding to α and $\eta\backslash\alpha$ from μ. Finally denote $R_{(\alpha,\alpha)}$ the matrix resulting from the extraction of the rows and columns both corresponding to α of Σ^{-1} and denote $S_{(\alpha,\eta\backslash\alpha)}$ the one resulting from the extraction of the rows corresponding to α and columns corresponding to $\eta\backslash\alpha$ of Σ^{-1}. Using Lemma 6 (and reordering indices if necessary) it is straightforward that $g_p(\alpha)$ is proportional to the density of a Gaussian distribution

$$\mathcal{N}(\mu_\alpha - R^{-1}_{(\alpha,\alpha)}S_{(\alpha,\eta\backslash\alpha)}(\eta\backslash\alpha - \mu_{\eta\backslash\alpha}), R^{-1}_{(\alpha,\alpha)})$$

2. Let us now consider the likelihood contribution. Using exactly the same notations that we used for the full conditional distribution of α for the algorithm associated to the non-informative approach we immediately find that $g_L(\alpha)$ is proportional to the density of a Gaussian distribution

$$\mathcal{N}([M'_\alpha M_\alpha]^{-1}M'_\alpha(y - Z_\alpha), \sigma^2 M'_\alpha M_\alpha)$$

just as in (8).

3. With the help of Lemma 6 and using the two results above, we can now deduce the posterior conditional distribution of α and obtain the Gaussian distribution

$$\alpha|\theta_*\backslash\alpha, y, \mathcal{D} \sim \mathcal{N}(\mu^\alpha, \Sigma^\alpha)$$

where

$$\begin{pmatrix} \mu^\alpha \\ \Sigma^\alpha \end{pmatrix} = \begin{pmatrix} \mu_\alpha - R^{-1}_{(\alpha,\alpha)}S_{(\alpha,\eta\backslash\alpha)}(\eta\backslash\alpha - \mu_{\eta\backslash\alpha}) \\ R^{-1}_{(\alpha,\alpha)} \end{pmatrix} * \begin{pmatrix} [M'_\alpha M_\alpha]^{-1}M'_\alpha(y - Z_\alpha) \\ \sigma^2 M'_\alpha M_\alpha \end{pmatrix}.$$

Full conditional distribution of β. Using similar arguments, we obtain the full conditional distribution of β. Namely, keeping the notation introduced to derive (9), and combining the prior and the likelihood contributions together with Lemma 6 we obtain the truncated Gaussian distribution

$$\beta|\theta_*\backslash\beta, y, \mathcal{D} \sim \mathcal{N}\left(\mu^\beta, \Sigma^\beta, B^{d_\beta}_+(0,1)\right)$$

where

$$\begin{pmatrix} \mu^\beta \\ \Sigma^\beta \end{pmatrix} = \begin{pmatrix} \mu_\beta - R^{-1}_{(\beta,\beta)}S_{(\beta,\eta\backslash\beta)}(\eta\backslash\beta - \mu_{\eta\backslash\beta}) \\ R^{-1}_{(\beta,\beta)} \end{pmatrix} * \begin{pmatrix} [M'_\beta M_\beta]^{-1}M'_\beta(y - Z_\beta) \\ \sigma^2 M'_\beta M_\beta \end{pmatrix}.$$

Full conditional distribution of γ. Using once again similar arguments, we obtain the full conditional distribution of γ. Namely, keeping the notation introduced to derive (10), and combining the prior and the likelihood contributions together with Lemma 6 we obtain the Gaussian distribution

$$\gamma|\theta_*\backslash\gamma, y, \mathcal{D} \sim \mathcal{N}(\mu^\gamma, \Sigma^\gamma)$$

where

$$\begin{pmatrix} \mu^\gamma \\ \Sigma^\gamma \end{pmatrix} = \begin{pmatrix} \mu_\gamma - R^{-1}_{(\gamma,\gamma)}S_{(\gamma,\eta\backslash\gamma)}(\eta\backslash\gamma - \mu_{\eta\backslash\gamma}) \\ R^{-1}_{(\gamma,\gamma)} \end{pmatrix} * \begin{pmatrix} [M'_\gamma M_\gamma]^{-1}M'_\gamma(y - Z_\gamma) \\ \sigma^2 M'_\gamma M_\gamma \end{pmatrix}.$$

Full conditional distribution of k. Using the definition of the hierarchical prior and Lemma 6 we immediately get

$$k|\theta_*\backslash k, y, \mathcal{D} \sim \mathcal{N}\left(\mu^k, \Sigma^k\right)$$

where

$$\begin{pmatrix} \mu^k \\ \Sigma^k \end{pmatrix} = \begin{pmatrix} q\,(1,\ldots,1)' \\ r^{-1} I_d \end{pmatrix} * \begin{pmatrix} (M^{\mathcal{A}})^{-1}\eta \\ l^{-1}\{(M^{\mathcal{A}})^{-1}\Sigma^{\mathcal{A}}(M^{\mathcal{A}})^{-1}\} \end{pmatrix}.$$

Full conditional distribution of l, q, r and σ^2). No calculations are required, as we respectively identify a gamma distribution, a Gaussian distribution, a gamma distribution, and an inverse-gamma distribution from (12).

5 Simulated data

We simulated the data using R (see R Development Core Team, 2008, for documentation) for ease of use and coded the algorithms presented in the earlier sections in language C (to benefit from short execution times). For any estimation (posterior mean and variance) on a dataset (be it $\mathcal{A}$ or $\mathcal{B}$), the MCMC algorithms would typically run for 500,000 iterations after a small burn-in period.

5.1 Comparing the hierarchical and the non-informative approaches

Predictive distribution. The Bayesian framework allows us to compute so-called predictive distributions, i.e. the distributions of future observations given past observations. Given a prior distribution $\pi(\theta)$ and the corresponding posterior distribution $\pi(\theta|y, \mathcal{D})$ related to the past observations $y = (y_1, \ldots, y_N)$ and data $\mathcal{D} = [\mathcal{D}_1, \ldots, \mathcal{D}_N]$, the predictive distribution for the future observation y_{N+k}, given data $\mathcal{D}_{N+k}$ is defined as

$$g(y_{N+k}|\mathcal{D}_{N+k}, y, \mathcal{D}) := \int f(y_{N+k}|\theta, \mathcal{D}_{N+k})\pi(\theta|y, \mathcal{D})\,\mathrm{d}\theta,$$

and the optimal prediction for the L^2 risk is then :

$$\widehat{y}_{N+k} := \mathbb{E}^{\pi}[y_{N+k}|\mathcal{D}_{N+k}, y, \mathcal{D}] \tag{13}$$

$$= \int y_{N+k} \cdot g(y_{N+k}|\mathcal{D}_{N+k}, y, \mathcal{D})\,\mathrm{d}y_{N+k}. \tag{14}$$

The comparison criterion. To assess the quality of the estimation of the model with our hierarchical prior with regard to the estimation of the model with the non-informative prior, we compare both results based on the quality of the predictions. Let y_{N+1} be the next upcoming observation, corresponding to data $\mathcal{D}_{N+1}$ and observe now that the prediction error can be written as

$$y_{N+1} - \widehat{y}_{N+1} = [y_{N+1} - \mu(\eta_0|\mathcal{D}_{N+1})] + [\mu(\eta_0|\mathcal{D}_{N+1}) - \widehat{y}_{N+1}],$$

which expresses the prediction error as a sum of a noise $y_{N+1} - \mu(\eta_0|\mathcal{D}_{N+1})$ (whose theoretical distribution is $\mathcal{N}(0, \sigma^2)$) and a bias which can be seen as an estimation error over the prediction $\mu(\eta_0|\mathcal{D}_{N+1}) - \widehat{y}_{N+1}$. We focus solely on the second part, since the first part (the noise) is unavoidable in real situation. Given that we want to validate our model on simulated data, the quantity $\mu(\eta_0|\mathcal{D}_{N+1}) - \widehat{y}_{N+1}$ is indeed accessible here whereas it would not be in real situation.

We thus choose to consider the quadratic distance between the real and the predicted model over a year as our quality criterion for a model, i.e. :

$$\sqrt{\frac{1}{365}\sum_{i=1}^{365}[\mu(\eta_0|\mathcal{D}_{N+i}) - \widehat{y}_{N+i}]^2}. \tag{15}$$

5.2 Construction of simulated datasets

Both datasets $\mathcal{A}$ and $\mathcal{B}$ were simulated according to the model (1) given on page 2 with $d_{11} = 4$ (4 frequencies used for the truncated Fourier series). The calendars and the partitions used for $\mathcal{A}$ and $\mathcal{B}$ were designed to include 7 daytypes ($d_2 = 7$, one daytype for each day of the week), but did not include any special days such as bankholidays. They also included 2 offsets ($d_{12} = 2$) to simulate the daylight saving time effect. In the end we thus had $d_\alpha = 4 \times 2 + 2 = 10$ and $d_\beta = 6$ i.e. $d = 19$ using the expression of the general model (2) given on page 4.

Dataset $\mathcal{A}$. We simulated 4 years of daily data for $\mathcal{A}$ with the following parameters :

$$\begin{aligned}
\alpha^{\mathcal{A}} &= (27, 7, -3, 1, 5, -1, 4, 0.5, 490, 495), \\
\beta^{\mathcal{A}} &= (0.13, 0.15, 0.16, 0.16, 0.16, 0.13), \\
\gamma^{\mathcal{A}} &= -3, \\
u^{\mathcal{A}} &= 14, \\
\sigma^{\mathcal{A}} &= 2.
\end{aligned}$$

These values were chosen to approximately mimic the typical electricity load of France up to a scaling factor. The temperatures we used for the estimation over $\mathcal{A}$ are those measured from September 1996 to August 2000 at 10:00AM.

Dataset $\mathcal{B}$. We simulated 1 year of daily data for $\mathcal{B}$ with the following parameters :

$$\begin{aligned}
&\alpha_i^{\mathcal{B}} = k_\alpha \times \alpha_i^{\mathcal{A}}, && && \forall i = 1, \ldots, d_\alpha \\
&\beta_1^{\mathcal{B}} = k_\beta \times \beta_1^{\mathcal{A}}, && \beta_j^{\mathcal{B}} = \beta_j^{\mathcal{A}}, && \forall j = 2, \ldots, d_\beta \\
&\gamma^{\mathcal{B}} = k_\gamma \times \gamma^{\mathcal{A}}, \\
&u^{\mathcal{B}} = k_u \times u^{\mathcal{A}}, \\
&\sigma^{\mathcal{B}} = 2,
\end{aligned}$$

where the coordinates of the true hyperparameters k were allowed to vary around 1. The temperatures we used for the estimation over $\mathcal{B}$ are those measured from September 2000 to August 2001 at 10:00AM.

We also simulated an extra year of daily data $\mathcal{B}$ for prediction, with the same parameters but with so-called normal temperatures, meaning that for each day of this extra year the temperature is the mean of all the past temperatures at the same time of the year. We made such a choice to try and suppress any dependency between our simulated results and the chosen temperature for this fictive year of prediction, since we did not want to bias our results due to a rigorous winter or an excessively hot summer.

5.3 Results

We chose to use vague priors (i.e. proper distributions with large variances) for the uppermost layers of our hierarchical prior, and thus decided to use the following values :

$$\begin{aligned}
\sigma_q &= 10^2, \\
a_r = b_r &= 10^{-6}, \\
a_l = b_l &= 10^{-3}.
\end{aligned}$$

A study of the Bayesian hierarchical model's sensitivity to these values showed that changing these hyperparameters to achieve prior variances of greater magnitudes hardly influenced the posterior results (means and variances) at all. This is why we decided to stick to these values for the remainder of our experimentations.

Estimation. We benchmarked the Bayesian model with its hierarchical prior against its original non-informative prior counterpart for different choices of true hyperparameters k over 300

replications (data being simulated anew for each replication), i.e. we simulated many different datasets $\mathcal{B}$ looking more or less similar to $\mathcal{A}$ and applied our method on them. Figures 3 and 5 show the posterior error of η (posterior mean minus the true value) and the posterior standard marginal deviations of η, based on 300 replications that correspond to the case where $k_\alpha = k_\beta = k_\gamma = k_u = 1$ i.e $\eta_\mathcal{A} = \eta_\mathcal{B}$ for both the informative (leftmost) and non-informative (rightmost) method. Marginal confidence interval for the posterior means are much smaller when using the hierarchical prior (most of them hitting the true value). The marginal posterior standard deviations are also reduced when using the hierarchical approach too.

When the situation is far from being as ideal as the one mentionned above, the hierarchical approach still shows improvement over the non-informative approach but to a lesser extent. Figures 4 and 6 show that the estimations of some of the parameters of the model are improved with the addition of the prior information (α and u) while some are not (β and γ) in the case where $k_\beta = k_u = 1$ and $k_\alpha = k_\gamma = 0.5$. Situations such as $k_\alpha = k_\gamma = k_u = 1$ and $k_\beta = 0.5$ or $k_\alpha = k_\gamma = k_\beta = 1$ and $k_u = 0.5$ were studied too and yielded very similar results i.e. lesser improvements on the estimations of some parameters only. Note that when some coordinates of k are valued to 0.5 while some are valued to 1, the "similarity" between $\mathcal{A}$ and $\mathcal{B}$ is very weak. The strength or weakness of the similarity between $\mathcal{A}$ and $\mathcal{B}$ cannot be diagnosed directly from the posterior mean of k itself but we will see that the estimations of the hyperparameters q and r may provide a partial answer to this problem.

We also estimated the hyperparameters (see section 4 for the definitions of hyperparameters k, l, r) when the hierarchical prior was used. Let us first study the hyperparameter k. Its coordinates seem correctly estimated for the ideal situation where $k_\alpha = k_\beta = k_\gamma = k_u = 1$ as illustrated in Figure 7 which shows the posterior error of k. When $k_\beta = k_u = 1$ and $k_\alpha = k_\gamma = 0.5$, the estimations obtained are of lesser quality as demonstrated in Figure 8 : most of the seasonal similarity coefficients appear to be biased (while the posterior standard deviation on each coordinate, not shown here, are greater than in the ideal situation). These estimations may thus be used to quantify the closeness of the two datasets.

The estimation of the hyperparameter l itself does not seem to provide a lot of information about the data : during our simulations, its mean value exhibited a lot of variability around the same value over the 300 replications for each of the five simulated scenarios and no reasonable conclusion could be drawn from it.

On the other hand, the estimation of the hyperparameter q does reveal a bit of information about the two datasets $\mathcal{A}$ and $\mathcal{B}$. It is the mean of the coordinates of k on the real axis, as can be seen in the definition of the hierarchical prior in (11) on page 11. However its use remains somewhat limited in the sense that the parameters β of the two datasets are most often very close (meaning the coordinates of k that correspond to them is likely close to 1) while other parameters may vary greatly. Hence even though q provides information about the similarity between $\mathcal{A}$ and $\mathcal{B}$, it cannot be interpreted alone and has to be considered jointly with r. Figure 9 shows the evolution of the posterior mean of q as $k_\alpha = k_u$ ranges over [0.5, 1].

The estimation of the hyperparameter r (inverse-variance of the prior distribution on k, see (11) again) does in fact reveal some information about the two datasets too. It is a measure of dispersion of k around q, in the sense that the (higher it is, the closer to q the coordinates of k should be. Just like q is the mean of the coordinates of k, r is in fact their inverse-variance. Figure 10 shows a clear decline when $k_\alpha = k_u$ moves away from the ideal value 1 i.e. when the similarity between the datasets $\mathcal{A}$ and $\mathcal{B}$ decrease from strong to weak.

As we previously stated, the similarity between the two datasets has to be assessed simultaneously with q and r and not q only : the mean q could be close to 1, possibly hinting at a perfect similarity between the two datasets, while the variance $1/r$ could be great which would then indicate huge differences between the two estimated sets of parameters for the two datasets.

Prediction. We compared the hierarchical and the non-informative models using our comparison criterion defined in (15) and computing the ratio betwen the two models for different values of k_α and k_γ, k_β and k_u being both set to 1. Figure 11 shows the results we obtained for k_α and k_γ simultaneously set to the values $1, 0.95, 0.90, 0.80$ and 0.50. Note that since the results appeared to be approximately symmetric with regard to 1 (i.e. for values $1, 1.05, 1.10, 1.20$ and 1.50), we

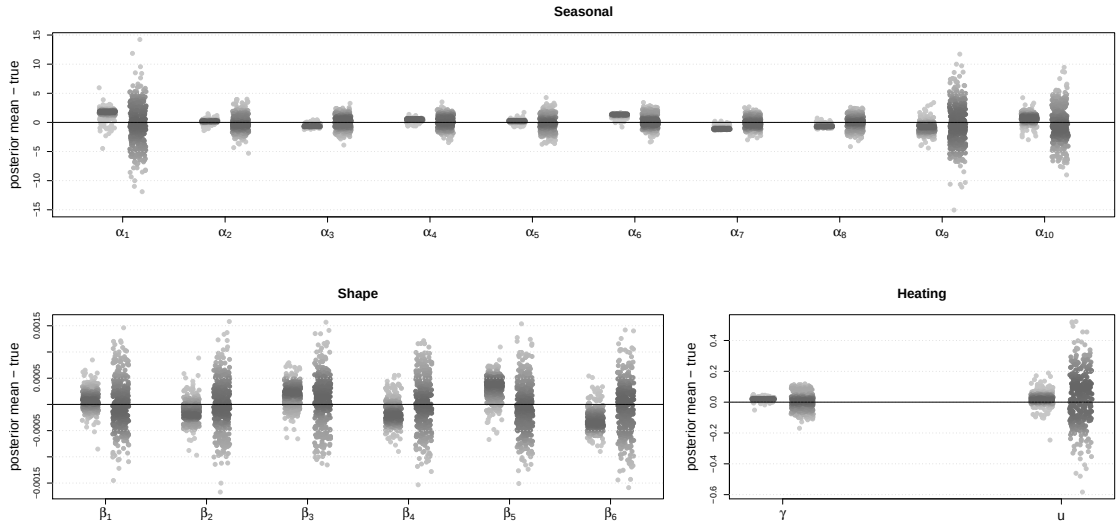

Figure 3: The posterior error (posterior mean minus true value) of α (seasonal parameters), β (shape parameters), and γ and u (heating parameters), based on 300 replications. Leftmost replications correspond to the hierarchical method while the rightmost replications correspond to the non-informative method. Here $k_\alpha = k_\beta = k_\gamma = k_u = 1$.

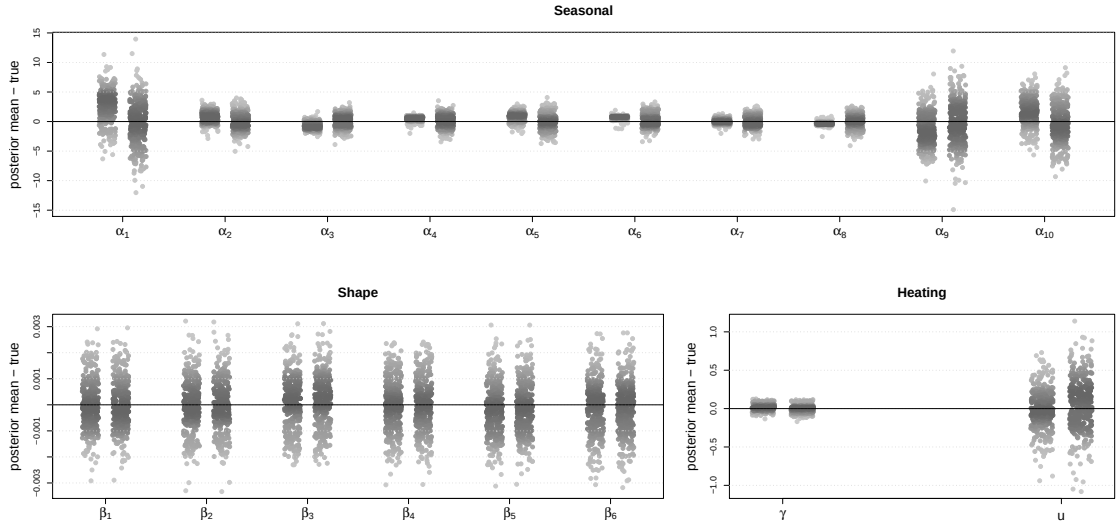

Figure 4: Same caption as in Figure 3 except $k_\beta = k_u = 1$ and $k_\alpha = k_\gamma = 0.5$.

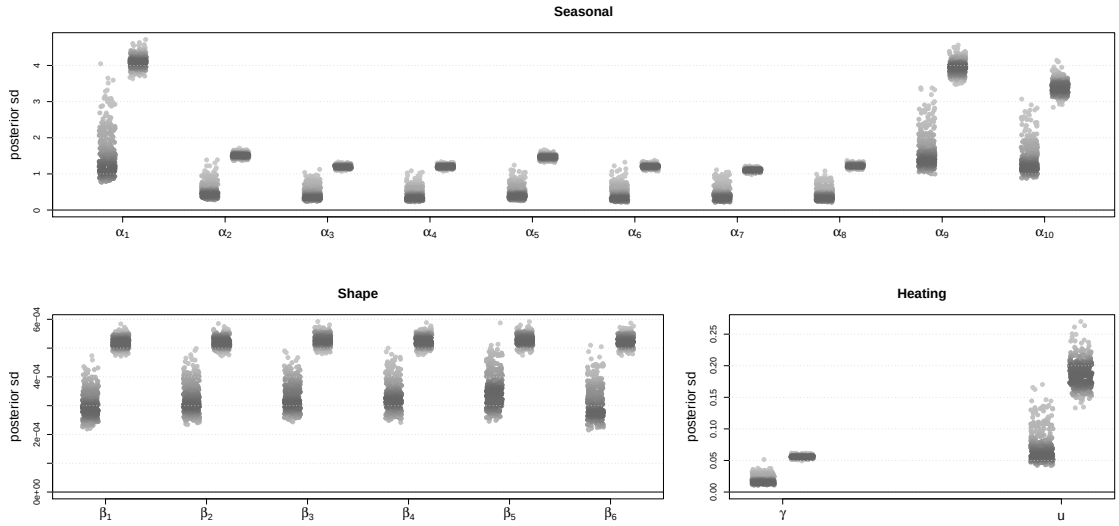

Figure 5: The posterior standard deviation of α (seasonal parameters), β (shape parameters), and γ and u (heating parameters), based on 300 replications. Leftmost replications correspond to the hierarchical method while the rightmost replications correspond to the non-informative method. Here $k_\alpha = k_\beta = k_\gamma = k_u = 1$.

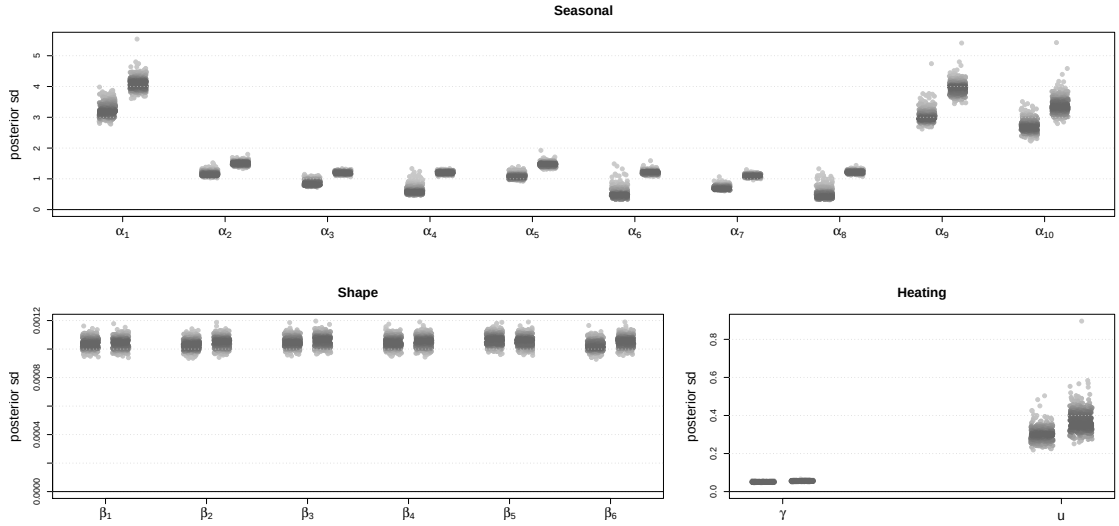

Figure 6: Same caption as in Figure 5 except $k_\beta = k_u = 1$ and $k_\alpha = k_\gamma = 0.5$.

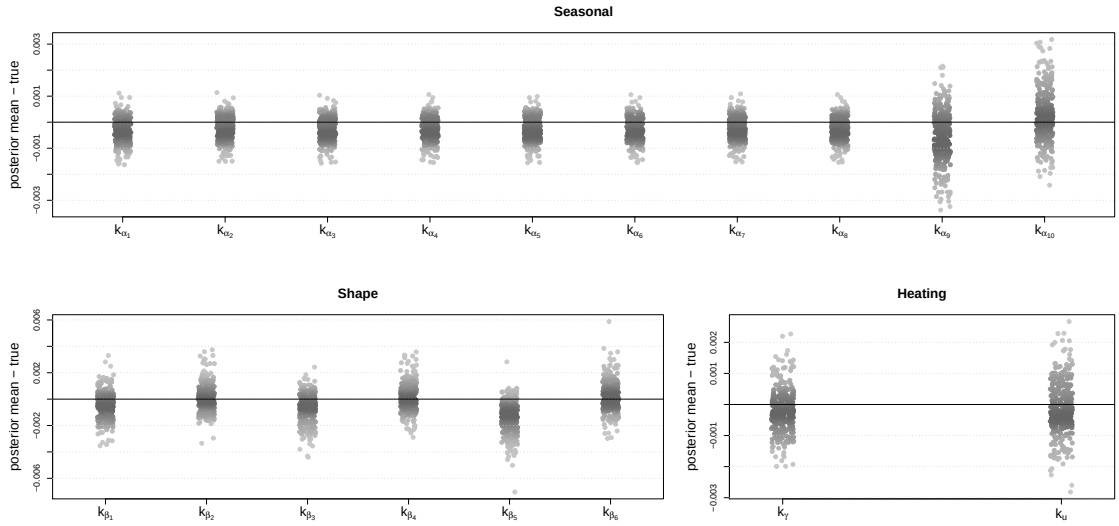

Figure 7: The posterior error (posterior mean minus true value) of k_α (seasonal parameters), k_β (shape parameters), and k_γ and k_u (heating parameters), based on 300 replications. Leftmost replications correspond to the hierarchical method while the rightmost replications correspond to the non-informative method. Here $k_\alpha = k_\beta = k_\gamma = k_u = 1$.

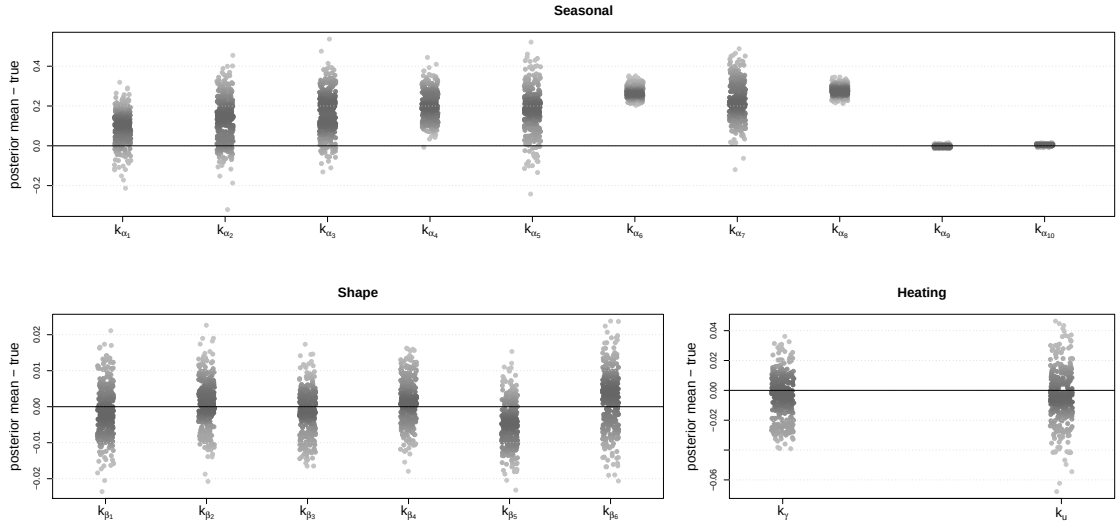

Figure 8: Same caption as in Figure 7 except $k_\beta = k_u = 1$ and $k_\alpha = k_\gamma = 0.5$.

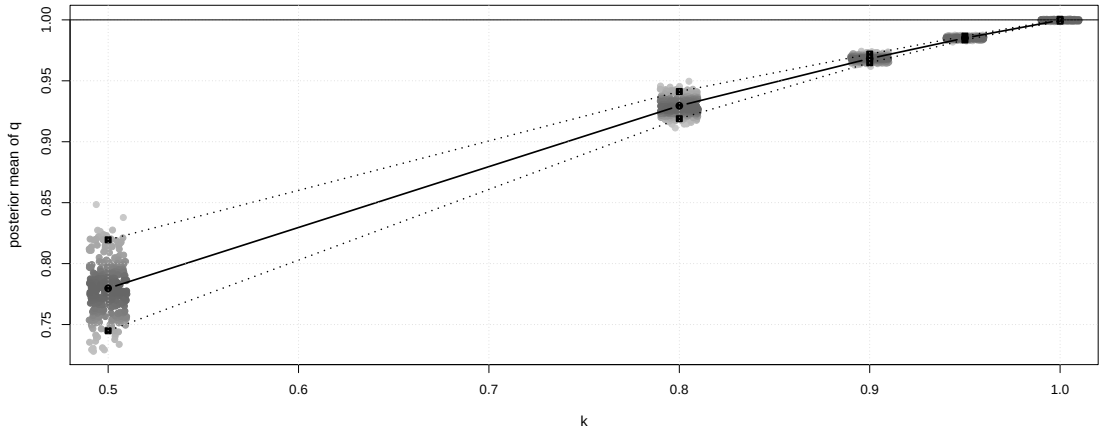

Figure 9: In grey : posterior mean of q for the hierarchical prior (abscissas have been jittered a bit to prevent overlapping, and different shades of grey are used to indicate the level of the estimated density). 300 replications for each value of $k_\alpha = k_\gamma$ tested. In black : the circles correspond to the averages, while the squares correspond to the 5% and 95% empirical quantiles. Here $k_\beta = k_u = 1$.

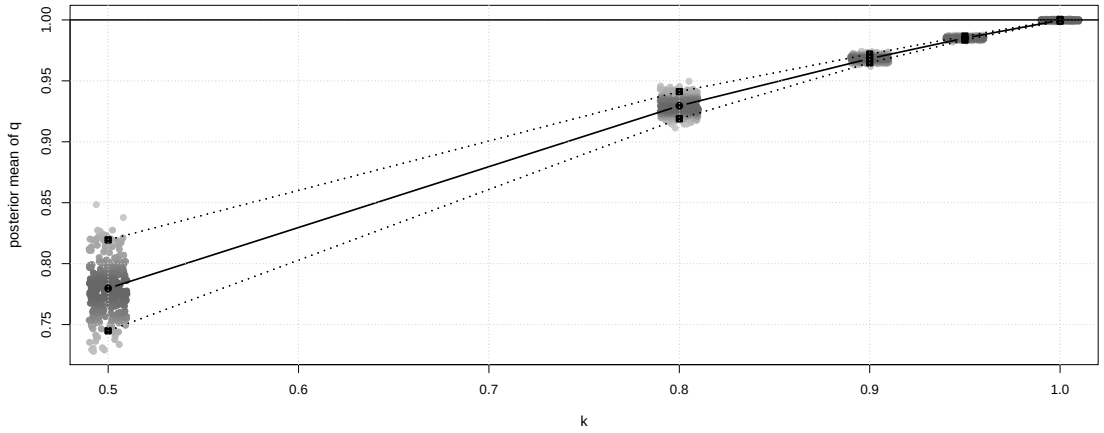

Figure 10: Same caption as in Figure 9 but for r (in log-scale).

only include one side of the graph in the present article.

On average, the Bayesian hierarchical model is a clear improvement over the Bayesian non-informative one, its performances being maximised when the parameters $\eta^{\mathcal{A}}$ and $\eta^{\mathcal{B}}$ are identical (which is the ideal situation). The performances in prediction are obviously somewhat weakened when the difference between the parameters $\eta^{\mathcal{A}}$ and $\eta^{\mathcal{B}}$ grows greater, but the use of the hierarchical model still leads to an average improvement of 15% over the non-informative model, as can be seen on Figure 11. The results obtained when k_β or k_u are varying while the other coordinates of k are fixed to 1 were very similar (see Figure 12).

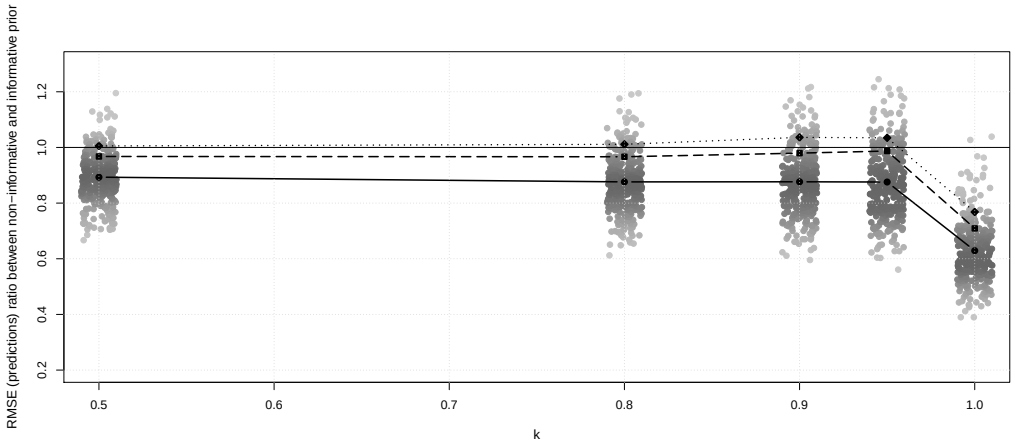

Figure 11: In grey : ratio between error predictions for the hierarchical and the non-informative approach (abscissas have been jittered a bit to prevent overlapping, and different shades of grey are used to indicate the level of the estimated density). 300 replications for each value of $k_\alpha = k_\gamma$ tested. In black : the circles correspond to the averages, while the squares and diamonds correspond to the 80% and 90% empirical quantiles of these ratios. Here $k_\beta = k_u = 1$.

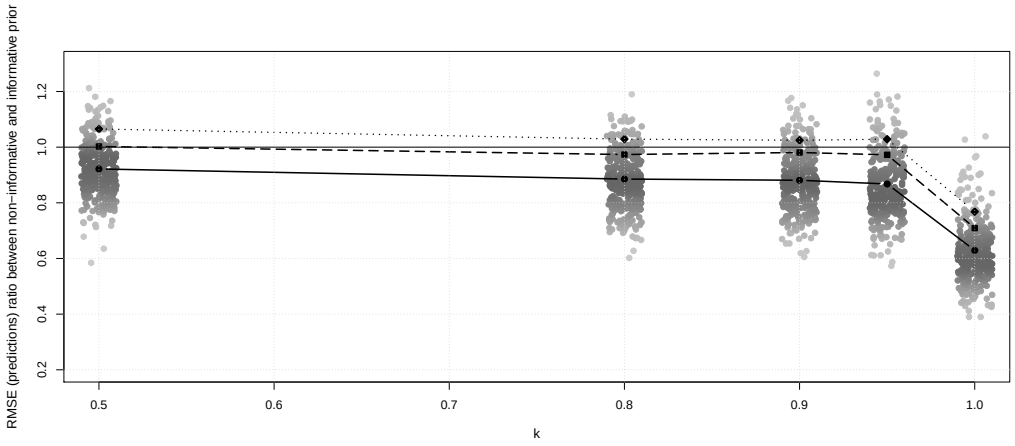

Figure 12: Same caption as in Figure 12 except here k_u varies and $k_\alpha = k_\beta = k_\gamma = 1$.

6 Applications

The dataset we used for $\mathcal{A}$ corresponds to a subpopulation from France frequently referred to as "non-metered" because their electricity consumption is not directly observed but instead deduced as the difference between the overall electricity consumption and the consumption of the "metered" population. We tested our method on two different populations $\mathcal{B} : \mathcal{B}_1$ which is a subpopulation from $\mathcal{A}$ and $\mathcal{B}_2$ which roughly covers the same people that $\mathcal{A}$ does. The sizes (in days) of the datasets are given in the Table 1 below.

To use our models on the datasets, we kept only one l oad value per day (the results shown hereafter were obtained for the load at 10:00AM) and trimmed the datasets so as to keep only days where the temperature would not exceed 18°C since our model does not take any cooling effect into account (symmetric to the heating effect, the cooling effect remains somewhat less important than the heating effect but we did not want to bias our estimations from the start with data known not to correspond to the chosen models).

$\mathcal{A}$	$\mathcal{B}_1$	$\mathcal{B}_2$
833	207	151

Table 1: Sizes of the real datasets (in days).

We kept the last 30 days of each $\mathcal{B}$ out of the estimation datasets and assessed the model quality over the predictions for those 30 days. It might seem an arbitrary choice and it is indeed, but the important lack of data prevented us from keeping 365 days as we previously did during the simulations. The procedure is similar in spirit to that developped for the simulations, but the results obtained in this section might be dependant on the temperature of these days, or their position in the calendar, while we did our best to avoid such a thing in the simulations. Restricting the prediction period to such a tiny time window might thus weaken somewhat the robustness of our method, but we nonetheless decided to show the performances we obtained on the real data in this paper.

6.1 Results on the large dataset $\mathcal{A}$

Using the non-informative prior over dataset $\mathcal{A}$ we are able to retrieve estimated predictive densities for future observations (see Figure 13) or alternatively we can estimate the quantiles of each of these densities to define credibility regions around the predictive mean. Most of the true observations lay well within the boundaries of the 95% credibility intervals of the predictions as can be seen on Figure 14.

6.2 Results on the small datasets $\mathcal{B}$

The estimation and prediction errors we obtained for the non-informative and the hierarchical methods on the two datasets $\mathcal{B}$ considered here are given in Table 2 below. While slightly degrading the quality of the fit on the estimation part compared to the non-informative approach, the hierarchical method vastly improves the quality of the predictions, reaching over 50% reduction for the root mean square error (RMSE) measure of accuracy.

The hierarchical prior allows us to retrieve information about the similarity between datasets $\mathcal{A}$ and $\mathcal{B}$ via the estimation of the posterior densities of the hyperparameters. The estimations of the posterior marginal distributions of k are presented on Figure 15 for both $\mathcal{B}_1$ and $\mathcal{B}_2$ and show how these datasets differ.

While the coordinates of k related to β seems to lie around 1, the rest of these coordinates do not concentrate around 1 for the dataset $\mathcal{B}_1$ as can be seen on the figures provided : the gaps ω_j of the model EVENTAIL defined in (1) are clearly centered around 0.5 while the rest of the coefficients linger somewhere around 0.7 or 0.8. Unlike $\mathcal{B}_1$, it seems $\mathcal{B}_2$ shares a lot of common features with $\mathcal{A}$

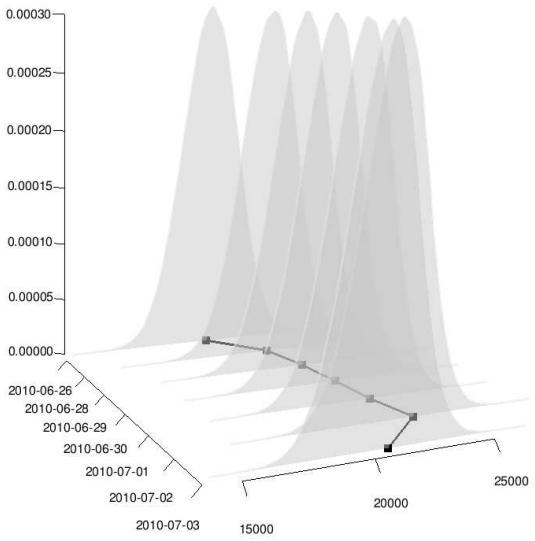

Figure 13: Estimated predictive densities for a few future days of the dataset $\mathcal{A}$. Future observed values are linked together with a black line.

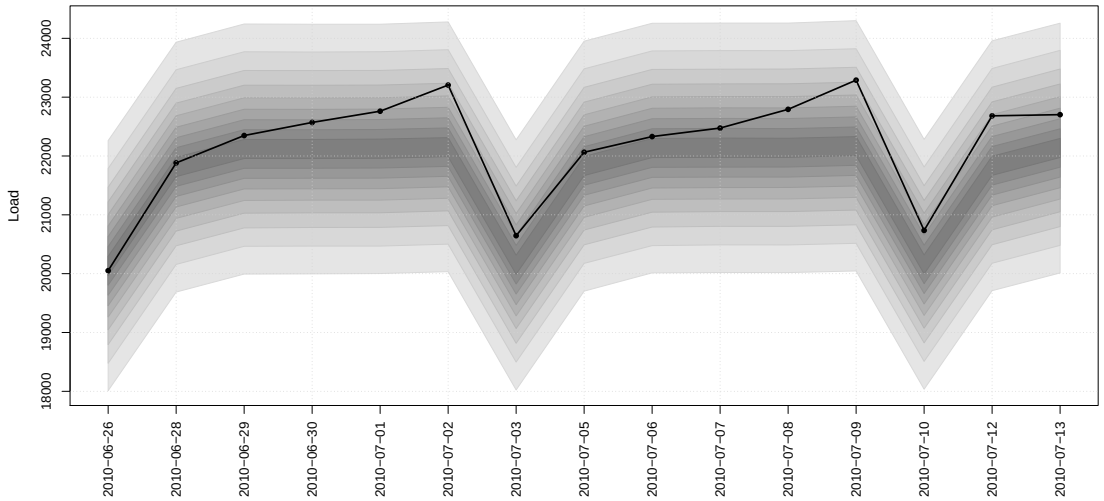

Figure 14: Estimated credibility regions for predictive densities for a few future days of the dataset $\mathcal{A}$. Future observed values are linked together with a blackline. The quantiles are drawn and linked in increasing shades of grey from 45% to 5% and from 55% to 95%.

	non-informative	hierarchical	comparison
$\mathcal{B}_1$			
RMSE est.	775.93	786.97	+1.42%
RMSE pred.	1863.25	894.00	−52.01%
MAPE est.	4.00	3.93	−0.07
MAPE pred.	19.37	9.30	−10.07
$\mathcal{B}_2$			
RMSE est.	1127.60	1202.32	+6.62%
RMSE pred.	2286.42	1339.14	−41.83%
MAPE est.	2.82	2.98	+0.15
MAPE pred.	8.65	3.48	−5.17

Table 2: Results for the dataset $\mathcal{B}_1$ and $\mathcal{B}_2$. RMSE is the "root mean square error" and MAPE is the "mean absolute percentage error". Both of these common measures of accuracy were computed on the estimation (est.) and prediction (pred.) parts of the two datasets.

: each marginal posterior density of k is peaked around 1 for $\mathcal{B}_2$ which indicates strong similarities. It is possible to derive credibility intervals on the mean values for each coordinate of k and these intervals are found to be smaller on $\mathcal{B}_2$ than they are on $\mathcal{B}_1$, as attested by the sharpness of the densities which are much more peaked on the former dataset than they are on the latter.

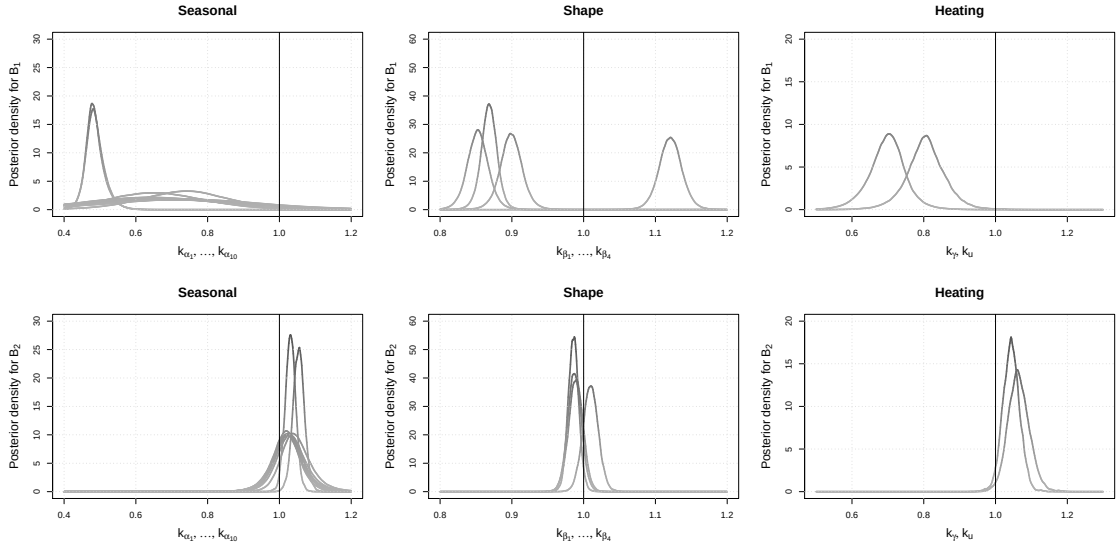

Figure 15: Estimated posterior marginal distributions of k for the hierarchical method and for both datasets $\mathcal{B}_1$ (upper row) and $\mathcal{B}_2$ (lower row). Coefficients corresponding to α, β and (γ, u) are shown on separate graphs.

The same conclusion can be drawn from the Table 3 in which we listed the estimated posterior means of l, q and r for $\mathcal{B}_1$ and $\mathcal{B}_2$: the estimated value of q (mean of all the coordinates of k) is closer to 1 on the second dataset than on the first and the estimated value of r (inverse-variance of all the coordinates of k) is greater too. These two hyperparameters can thus be used to quickly assess the strength of the similarity between the two datasets $\mathcal{A}$ and $\mathcal{B}$ while only a close study of the posterior marginal densities of k can reveal which coordinates are similar and which are not.

In fact the upper row of Figure 15 suggests that the specification of the hierarchical prior as

a mixture of normal distributions $\mathcal{N}(q_i, r_i)$ could possibly help in getting even better results on dataset $\mathcal{B}_1$, to help distinguish at least two groups for the coordinates of k using their means : the coordinates that are close to 1, and those that are not.

	$\mathcal{B}_1$	$\mathcal{B}_2$
l	19.17	128.60
q	0.73	1.02
r	24.48	795.16

Table 3: Estimated posterior mean of the hyperparameters l, q and r for both of the studied datasets. These estimations may serve as a summary of the studies : the similarity between $\mathcal{A}$ and $\mathcal{B}_2$ is found to be stronger than the one between $\mathcal{A}$ and $\mathcal{B}_1$ as the posterior mean of q (mean of the similarity coefficients k_i) and r (inverse-variance of the similarity coefficients k_i) indicate together.

References

Al-Zayer, J. and Al-Ibrahim, A. (1996). Modelling the impact of temperature on electricity consumption in the eastern province of saudi arabia. *Journal of Forecasting*, 15:97–106.

Bruhns, A., Deurveilher, G., and Roy, J. (2005). A non-linear regression model for mid-term load forecasting and improvements in seasonnality. *Proceedings of the 15th Power Systems Computation Conference 2005, Liege Belgium*.

Cottet, R. and Smith, M. (2003). Bayesian modeling and forecasting of intraday electricity load. *Journal of the American Statistical Association*, 98(464):839–849.

Gallant, A. (1975). Non-linear regression. *The American Statistician*, 29(2):73–81.

Gallant, A. and Fuller, W. (1973). Fitting segmented polynomial regression models whose join points have to be estimated. *Journal of the American Statistical Association*, 68(341):144–147.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Harrison, P. and Stevens, C. (1976). Bayesian forecasting. *Journal of the Royal Statistical Society*, 38(3):205–247.

Hinkley, D. (1971). Inference in two-phase regression. *Journal of the American Statistical Association*, 66(336):736–743.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). Winbugs - a bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325–337.

Marin, J.-M. and Robert, C. (2007). *Bayesian Core : A Practical Approach to Computational Bayesian Statistics*. Springer.

Marriott, J. and Spencer, N. (2001). A note on bayesian prediction from the regression model with informative priors. *Aust. N.Z. J. Stat.*, 43(4):473–480.

Minka, T. P. (1999). Bayesian linear regression. Technical report, 3594 Security Ticket Control.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robert, C. (1995). Simulation of truncated normal variables. *Statistics and Computing*, 5(2):121–125.

Roberts, G. and Rosenthal, J. (2001). Optimal scaling for various metropolis-hastings algorithms. *Statistical Science*, 16(4):351–367.

Seber, G. and Wild, C. (2003). *Nonlinear Regression*. Wiley.